

A Hierarchical Bayesian Framework for Constructing Sparsity-inducing Priors

Anthony Lee ^{*} Francois Caron [†] Arnaud Doucet [‡] Chris Holmes [§]

September 20, 2010

Abstract

Variable selection techniques have become increasingly popular amongst statisticians due to an increased number of regression and classification applications involving high-dimensional data where we expect some predictors to be unimportant. In this context, Bayesian variable selection techniques involving Markov chain Monte Carlo exploration of the posterior distribution over models can be prohibitively computationally expensive and so there has been attention paid to quasi-Bayesian approaches such as maximum *a posteriori* (MAP) estimation using priors that induce sparsity in such estimates. We focus on this latter approach, expanding on the hierarchies proposed to date to provide a Bayesian interpretation and generalization of state-of-the-art penalized optimization approaches and providing simultaneously a natural way to include prior information about parameters within this framework. We give examples of how to use this hierarchy to compute MAP estimates for linear and logistic regression as well as sparse precision-matrix estimates in Gaussian graphical models. In addition, an adaptive group lasso method is derived using the framework.

1 Introduction

There has been recent interest in sparse estimates for coefficients in regression problems, with this problem often termed variable selection in the literature. To this end, a variety of approaches have been proposed in both the statistics and signal processing literatures. Most of the computationally tractable approaches are the solutions of penalized optimization problems associated with regularization of the coefficients in likelihood optimization. Although not truly Bayesian approaches, often they can be interpreted as maximum *a posteriori* (MAP) estimates associated with the posterior density of the coefficients where the prior induces the regularization used in the optimization routine.

Denoting the coefficients by $\beta \in \mathbb{R}^p$, a popular family of these computing estimates as solutions to penalized optimization problems involving the log-likelihood of the data given β and ℓ_q

^{*}Oxford-Man Institute and Department of Statistics, University of Oxford, UK

[†]INRIA Bordeaux Sud-Ouest and Institut de Mathématiques de Bordeaux, University of Bordeaux, France

[‡]Institute of Statistical Mathematics, Japan and University of British Columbia, Department of Statistics and Department of Computer Science, Canada.

[§]Department of Statistics and Oxford-Man Institute, University of Oxford, UK

penalization on the coefficients with $0 \leq q \leq 1$. When q is in this range, the solutions are sparse for large enough values of multiplicative penalization weights. When $q \geq 1$, the penalization is additionally convex, a property that has made the choice of $q = 1$ particularly suitable when the log-likelihood is concave as this leads to a unique global maxima for the objective function.

Beginning with [1], it has become popular practice to use ℓ_1 -regularization on each component of β . However, use of identical penalization on each coefficient, e.g. $\lambda \sum_{j=1}^p |\beta_j|$ can lead to unacceptable bias in the resulting estimates [2], which has motivated use of sparsity-inducing non-convex penalties despite the increased difficulty in computing the resulting estimates. In particular, this has led to the adoption of “adaptive” methods [3, 4] in the statistics literature and iteratively reweighted methods [5, 6] in the signal processing literature.

We propose a hierarchical prior for β that amounts marginally to a sparsity-inducing, non-convex penalty in MAP estimation. Further, the specific hierarchy gives rise to an expectation-maximization (EM) algorithm [7] that is essentially an iteratively reweighted ℓ_q -minimization algorithm. In one case, the algorithm corresponds to the iteratively reweighted ℓ_1 -minimization algorithm and has been independently suggested in both [8] and [9]. Our hierarchical formulation of the prior, in contrast, allows users to incorporate prior information about different coefficients and allows flexibility in grouping variables together. For example, the framework gives immediately an adaptive version of the group lasso algorithm proposed in [10].

2 The hierarchical adaptive lasso (HAL)

We are interested in prior distributions for β in a general regression settings. Let $\mathbb{T}_k \stackrel{\text{def}}{=} \{1, \dots, k\}$. We are given n observations $\{y_i\}_{i=1}^n$ and associated with each observation a vector of covariates $\mathbf{x}_i \in \mathbb{R}^p$ for $i \in \mathbb{T}_n$. We assume that the conditional distribution of each y_i is independent given \mathbf{x}_i and has density $f(y|\mathbf{x}, \beta, \theta)$, where $\beta \in \mathbb{R}^p$ and $\theta \in \Theta$ parametrize the distribution of y conditional on \mathbf{x} . Defining $\mathbf{y} \stackrel{\text{def}}{=} (y_1, \dots, y_n)' \in \mathbb{R}^n$ and $\mathbf{X} \stackrel{\text{def}}{=} (\mathbf{x}'_1, \dots, \mathbf{x}'_n)' \in \mathbb{R}^{n \times p}$, the conditional distribution of all of the observations is then given by $f(\mathbf{y}|\mathbf{X}, \beta, \theta) \stackrel{\text{def}}{=} \prod_{i=1}^n f(y_i|\mathbf{x}_i, \beta, \theta)$. We are primarily interested in the parameter β and assume that each component β_j has special meaning when equal to 0.

While a Bayesian approach would usually suggest approximating the posterior density

$$p(\beta|\mathbf{y}, \mathbf{X}, \theta) \propto f(\mathbf{y}|\mathbf{X}, \beta, \theta)p(\beta|\theta),$$

we focus here on MAP (point) estimates of β since these are computationally easier to compute, especially when $f(\mathbf{y}|\mathbf{X}, \beta, \theta)$ and $p(\beta|\theta)$ are concave, and their use is not uncommon when p and/or n are large. MAP estimates are computed by solving the optimization problem

$$\hat{\beta}_{MAP} = \arg \max_{\beta} f(\mathbf{y}|\mathbf{X}, \beta, \theta)p(\beta|\theta)$$

or, equivalently

$$\hat{\beta}_{MAP} = \arg \max_{\beta} \log f(\mathbf{y}|\mathbf{X}, \beta, \theta) + \log p(\beta|\theta)$$

The $\log p(\beta|\theta)$ can be thought of as a penalization term in optimizing the log-likelihood of the data $\log f(\mathbf{y}|\mathbf{X}, \beta, \theta)$.

2.1 Generalized t-distribution prior

We propose a hierarchical approach to constructing priors for β . At the lowest level, we give each element β_j of β an independent normal prior with mean 0 and variance σ_j^2 , ie. $p(\beta|\sigma_{1:p}^2) = \prod_{j=1}^p p(\beta_j|\sigma_j^2)$ where $\beta_j|\sigma_j^2 \sim N(0, \sigma_j^2)$. If we leave σ_j^2 for $j \in \mathbb{T}_p$ as hyperparameters, computing the resulting MAP estimate corresponds to ℓ_2 -penalized optimization of the log-likelihood.

If, instead, we model each σ_j^2 as being drawn from an exponential distribution with mean $2\tau_j^2$ we obtain a double-exponential distribution for β_j after σ_j^2 has been integrated out, ie.

$$p(\beta_j|\tau_j) = \frac{1}{2\tau_j} \exp\left(-\frac{|\beta_j|}{\tau_j}\right)$$

Computing the MAP estimate associated with this prior corresponds to ℓ_1 -penalized optimization and the solution itself is identical to the LASSO estimate when f is a multivariate Gaussian with mean $\mathbf{X}\beta$. This prior has become popular in recent years for variable selection since it induces sparsity in $\hat{\beta}_{MAP}$ for small enough values of τ_j .

We propose adding another level of hierarchy to the prior by having separate random variables τ_j for each $j \in \mathbb{T}_p$ and placing inverse-gamma priors on each τ_j . Indeed, if we let $\tau_j \sim IG(a_j, b_j)$ we obtain

$$p(\beta_j|a_j, b_j) = \frac{a_j}{2b_j} \left(\frac{|\beta_j|}{b_j} + 1 \right)^{-(a_j+1)} \quad (1)$$

after integrating out τ_j , which we call the hierarchical adaptive lasso (HAL) prior since one can compute MAP estimates using this prior with a type of adaptive lasso algorithm, as can be seen in Section 2.2. This is the density of a generalized t-distribution. Computing the MAP estimate associated with this prior corresponds to logarithmic penalization of the log-likelihood. From a Bayesian modelling perspective, the introduction of a distribution over the τ_j is a natural way to resolve the issue of believing that there are significant differences in the sizes of the coefficients of β that cannot be modelled as having come from a distribution with as thin tails as a Laplace distribution.

2.2 Computing MAP estimates

The optimization problem associated with the generalized t-distribution prior is not concave. However, one can find local modes of the posterior using the EM algorithm with the $\tau = \tau_{1:p}$ as latent variables. Indeed, each iteration of EM takes the form

$$\beta^{(t+1)} = \arg \max_{\beta} \log f(\mathbf{y}|\mathbf{X}, \beta, \theta) + \int \log[p(\beta|\tau)]p(\tau|\beta^{(t)}, a, b)d\tau$$

The conjugacy of the inverse-gamma distribution with respect to the Laplace distribution gives

$$\tau_j|\beta_j^{(t)}, a_j, b_j \sim IG(a_j + 1, b_j + |\beta_j|)$$

and with $p(\beta|\tau) = \prod_{j=1}^p p(\beta_j|\tau_j) = \prod_{j=1}^p 1/(2\tau_j) \exp(-|\beta_j|/\tau_j)$ yields

$$\beta^{(t+1)} = \arg \max_{\beta} \log f(\mathbf{y}|\mathbf{X}, \beta, \theta) - \sum_{j=1}^p |\beta_j| \int \frac{1}{\tau_j} p(\tau_j|\beta_j^{(t)}, a_j, b_j) d\tau_j$$

where the expectation of $1/\tau_j$ given $\tau_j \sim IG(a_j + 1, b_j + |\beta_j^{(t)}|)$ is $(a_j + 1)/(b_j + |\beta_j^{(t)}|)$.

As such, one can find a local mode of the posterior $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \theta)$ by starting at some point $\boldsymbol{\beta}^{(0)}$ and then iteratively solving

$$\boldsymbol{\beta}^{(t+1)} = \arg \max_{\boldsymbol{\beta}} \log f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \theta) - \sum_{j=1}^p w_j^{(t)} |\beta_j|$$

where

$$w_j^{(t)} = \frac{a_j + 1}{b_j + |\beta_j^{(t)}|}$$

It is clear that for large enough values of a_j and small enough values of b_j that the MAP estimates obtained by the EM algorithm are sparse. In fact, any posterior mode with this prior corresponds to a weighted lasso solution, which is sparse when the penalization through $\{(a_j, b_j)\}_{j=1}^p$ is large enough.

2.2.1 Oracle properties

In the penalized optimization literature, some methods are justified at least partially by their possession of the oracle property: that for appropriate parameter choices, the method performs just as well as an oracle procedure in terms of selecting the correct variables and estimating the nonzero coefficients asymptotically in n . Using the HAL prior in Theorem 5 of [4] gives us the oracle property if $a_j \rightarrow \infty$ and $n^{-1/2}a_j \rightarrow 0$ as $n \rightarrow \infty$. It is worth remarking that this property requires our prior on $\boldsymbol{\beta}$ to depend on the number of observations, which is atypical in Bayesian inference. Intuitively, a_j needs to increase as n increases to ensure that the solution remains sparse whilst it cannot increase too quickly or consistency is lost. As pointed out in [2], this trade off is impossible to accomplish with the LASSO.

2.3 Generalizations and extensions

2.3.1 Exponential power family

One can model β_j more generally as coming from an exponential power distribution instead of a Laplace distribution. In this case, we can write

$$p(\beta_j|\eta_j, q) = \frac{1}{2\eta_j^{1/q}\Gamma(1 + 1/q)} \exp\left(-\frac{|\beta_j|^q}{\eta_j}\right)$$

With an inverse-gamma prior on η_j , which enjoys conjugacy with respect to the exponential power distribution, we obtain

$$p(\beta_j|a_j, b_j, q) = \frac{\Gamma(a_j + 1/q)}{2\Gamma(a_j)\Gamma(1 + 1/q)b_j^{1/q}} \left(\frac{|\beta_j|^q}{b_j} + 1\right)^{-(a_j + 1/q)}$$

Use of this prior results in the same algorithm but with the weights given by

$$w_j^{(t)} = \frac{a_j + 1/q}{b_j + |\beta_j^{(t)}|^q}$$

The use of an exponential power prior can be motivated hierarchically as a scale mixture of normal distributions for $q \in [1, 2)$ [11] or as a scale mixture of uniform distributions for $q \in (1, \infty)$ [12]. For $q \in (0, 1)$ this distribution is still defined but it does not have the same interpretation as when $q \geq 1$ and additionally has a non-concave density which complicates computation of posterior modes. The choice $q = 2$ corresponds to a normal distribution and after marginalizing out η it gives a scaled t -distribution with $2a_j$ degrees of freedom and scale $\sqrt{b_j/a_j}$. This choice leads to a hierarchical adaptive ℓ_2 -regularized method that may be suitable for problems in which prediction instead of variable selection is more important.

Contour plots of the negative log density of the joint prior for two variables are given in Figure 1 and thresholding plots associated with the priors are given in Figure 2. The contour plots show graphically how the LASSO and HAL approaches give sparse solutions whilst the hierarchical adaptive ridge (HAR) prior, corresponding to $q = 2$, gives non-sparse solutions. The thresholding plots show that whilst the LASSO significantly biases even large coefficients, the HAL and HAR do not.

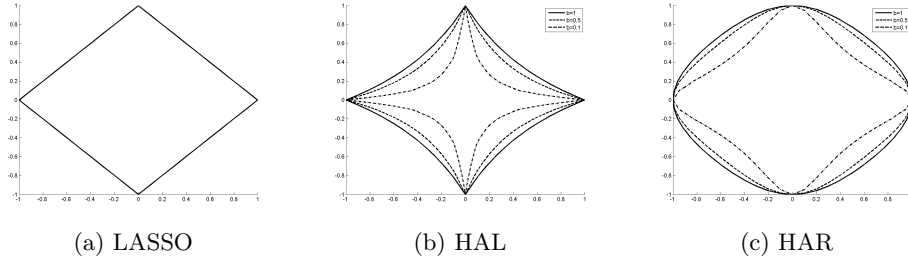


Figure 1: Two-dimensional contour plots of the penalties, ie. the negative log density, associated with the priors.

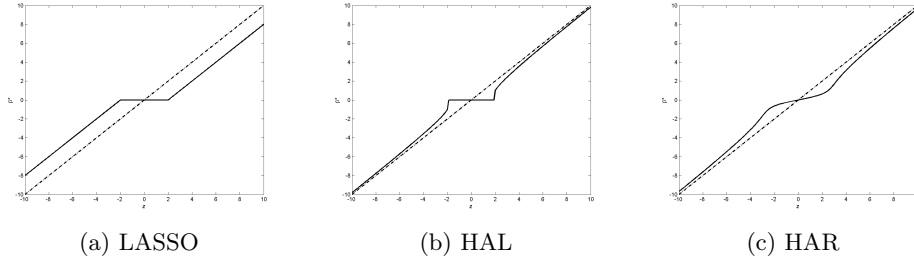


Figure 2: Threshold plots associated with the priors.

2.3.2 The hierarchical adaptive group lasso

The hierarchical framework allows us to group variables together by making them dependent on a shared variable higher up in the hierarchy. For example, letting $g : \mathbb{T}_p \rightarrow \mathbb{T}_K$ be a function mapping variables to one of K groups and n_i be the number of variables in group i we can use

the following model:

$$\begin{aligned}\beta_j | \sigma_{g(j)}^2 &\sim N(0, \sigma_{g(j)}^2), j \in \mathbb{T}_p \\ \sigma_i^2 | \tau_i &\sim G\left(\frac{n_i + 1}{2}, 2\tau_i^2\right), i \in \mathbb{T}_K \\ \tau_i | a_i, b_i &\sim IG(a_i, b_i), i \in \mathbb{T}_K\end{aligned}$$

With $G_i = \{j : g(j) = i\}$, this gives

$$p(\beta_{G_i} | \tau_i) = \frac{(2\tau_i)^{-n_i} \pi^{-(n_i-1)/2}}{\Gamma((n_i + 1)/2)} \exp\left(-\frac{\sqrt{\sum_{j \in G_i} |\beta_j|^2}}{\tau_i}\right)$$

and so $\tau_i | \beta_{G_i}, a_i, b_i \sim IG(a_i + n_i, b_i + \sqrt{\sum_{j \in G_i} |\beta_j|^2})$. The corresponding iterative procedure is then

$$\beta^{(t+1)} = \arg \max_{\beta} \log f(\mathbf{y} | \mathbf{X}, \beta, \theta) - \sum_{i=1}^K w_i^{(t+1)} \|\beta_{G_i}\|_2$$

where

$$w_i^{(t+1)} = \frac{a_i + n_i}{\|\beta_{G_i}^{(t)}\|_2 + b_i}$$

The marginal prior on β_{G_i} has the density

$$p(\beta_{G_i} | a_i, b_i) = \frac{(2b_i)^{-n_i} \pi^{-(n_i-1)/2} \Gamma(n_i + a_i)}{\Gamma((n_i + 1)/2) \Gamma(a_i)} \left(\frac{\|\beta_{G_i}^{(t)}\|_2}{b_i} + 1 \right)^{(-a_i - n_i)}$$

but this density is never evaluated in the EM algorithm.

A related problem to grouped variable selection is known as multi-task learning within the machine learning literature, where one wants to solve for $\theta \stackrel{\text{def}}{=} \{\beta^{(i)}\}_{i=1}^L$ in a variety of L related regression models. One approach is to solve the optimization problem

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \sum_{i=1}^L \log f_i(\mathbf{y}_i | \mathbf{X}_i, \beta^{(i)}) + \sum_{j=1}^p \lambda_j \|\beta_j\|_2$$

where $\beta_j \stackrel{\text{def}}{=} (\beta_j^{(1)}, \dots, \beta_j^{(L)}) \in \mathbb{R}^L$ [13]. This type of regularization can be derived using the same hierarchical prior used in the group lasso where the coefficients relating to the same covariate are ‘grouped’ together to promote sparsity across the individual β estimates, ie. a covariate is selected in all the related models or in none of the models. As such, an adaptive version of this multi-task learning approach follows the same form as the hierarchical adaptive group lasso.

2.3.3 Matrix priors

For the purpose of covariance matrix estimation, ℓ_1 -regularization has been used on entries of the precision matrix Ω of a Gaussian graphical model [14, 15]. This corresponds to MAP estimation using Laplace priors on each Ω_{ij} for $i \leq j$. We can incorporate this type of prior

within our framework by placing inverse Gamma priors on the scale parameters of each Laplace distribution. We have

$$p(\Omega_{ij}|\tau_{ij}) = \frac{1}{2\tau_{ij}} \exp\left(-\frac{|\Omega_{ij}|}{\tau_{ij}}\right)$$

with $p(\Omega|\tau) = \prod_{i=1}^p \prod_{j=i}^p p(\Omega_{ij}|\tau_{ij})$ and $\tau_{ij} \sim IG(a_{ij}, b_{ij})$. Note that in this formulation the prior on Ω is non-zero for non-positive-definite values. This allows us to specify

$$p(\tau|\Omega, A, B) = \prod_{i=1}^p \prod_{j=i}^p IG(\tau_{ij}; a_{ij} + 1, b_{ij} + |\Omega_{ij}|)$$

One can have the likelihood of observed data Y , $p(Y|\Omega)$ be zero if Ω is not symmetric positive-definite. In this case, the posterior and hence the MAP estimate are equivalent to the case where the prior takes the form

$$p(\Omega|A, B) = \frac{\mathbf{1}_{\mathcal{P}}(\Omega)p(\Omega|A, B)}{\int \mathbf{1}_{\mathcal{P}}(\Omega)p(\Omega|A, B)d\Omega}$$

since in both cases we have

$$p(\Omega|Y, A, B) = \frac{\mathbf{1}_{\mathcal{P}}(\Omega)p(Y|\Omega)p(\Omega|A, B)}{\int \mathbf{1}_{\mathcal{P}}(\Omega)p(Y|\Omega)p(\Omega|A, B)d\Omega}$$

where \mathcal{P} is the set of symmetric positive-definite matrices. Note, however, that the positive-definite prior cannot be used to derive the EM algorithm central to our methodology since $\tau|\Omega, A, B$ is no longer a product of inverse-gamma distributions.

2.3.4 The hierarchical lasso

In some cases, one might be interested in having $\eta_j = \eta$ for all $j \in \mathbb{T}_p$ with $\eta \sim IG(a, b)$. In this case, one obtains a prior on β of the form

$$p(\beta|a, b, q) = \frac{\Gamma(a + p/q)}{2^p \Gamma(a) \Gamma(1 + 1/q)^p b^{p/q}} \left(\frac{\sum_{j=1}^p |\beta_j|^q}{b} + 1 \right)^{-a-p/q} \quad (2)$$

which leads to the iterative procedure

$$\beta^{(t+1)} = \arg \max_{\beta} \log f(\mathbf{y}|\mathbf{X}, \beta, \theta) - w^{(t+1)} \sum_{j=1}^p |\beta_j|^q$$

where

$$w^{(t+1)} = \frac{a + p/q}{b + \sum_{j=1}^p |\beta_j^{(t)}|^q}$$

In fact, a more general prior can be constructed by considering groupings of the coefficients such that $\beta_j \sim EP(\eta_i, q)$ for all $j \in G_i$, where G_i is again the set of indices of coefficients in group i . A prior constructed in this fashion leads to the iterative procedure

$$\beta^{(t+1)} = \arg \max_{\beta} \log f(\mathbf{y}|\mathbf{X}, \beta, \theta) - \sum_{i=1}^K w_i^{(t+1)} \sum_{j \in G_i} |\beta_j|^q$$

where

$$w_i^{(t+1)} = \frac{a_i + n_i/q}{b_i + \sum_{j \in G_i} |\beta_j^{(t)}|^q}$$

2.3.5 Modifying the hierarchy

The above examples are only a subset of the possible modifications to the hierarchy that are possible. Indeed, one of the benefits of a hierarchical approach is that one can flexibly group variables via the sharing of random variables. Graphical models for the exponential family generalization and the grouped variable generalization are given in Figure 3 along with a discussion of their relationships to existing methods in Section 3.7.

2.4 Tuning the hyperparameters

Use of the proposed framework relies on appropriate settings of the hyperparameters. For distributions of non-negative Z with density

$$p(Z|\nu, b) = \frac{\nu - 1}{b} \left(\frac{Z}{b} + 1 \right)^{-\nu}$$

the moments of Z are given by

$$\mathbb{E}_p[Z^t] = \frac{b^t \Gamma(\nu - 1 - t) \Gamma(t + 1)}{\Gamma(\nu - 1)}$$

which allows one to pick hyperparameters that represent prior beliefs about the mean and variance of variables of interest, e.g. $|\beta_j|$ in the case of prior (1) or $(\sum_{j=1}^p |\beta_j|^q)^{1/q}$ in the case of prior (2).

Focusing on the hierarchical adaptive lasso prior, we note that in this case we have $\mathbb{E}[|\beta_j|] = b_j/(a_j - 1)$, for $a_j > 1$. An observation on a_j and b_j is that when one increases both values but keeps $\mathbb{E}[|\beta_j|]$ constant, the tendency for the iterative scheme to set β_j to zero is reduced since w_j is upper-bounded by $(a_j + 1)/b_j$. This observation could be used in a ‘tempered’ optimization scheme as discussed in Section 2.5, noting in particular that as $a_j \rightarrow \infty$, the prior approaches a Laplace distribution and so the posterior approaches unimodality.

2.5 Issues with MAP estimation

There are many criticisms of MAP estimates in a Bayesian framework. We motivate use of such estimates for primarily computational reasons, since Bayesian variable selection methods tend to be prohibitively expensive when dealing with large data sets. Beyond the obvious problem of summarizing the posterior distribution over models with a point estimate, one problem is that MAP estimates are not Bayes estimators but instead a limit of Bayes estimators under the 0-1 loss function. While important, this issue is not addressed here. A perhaps more fundamental issue is that MAP estimates are not invariant under reparametrization. This issue can be rectified by finding the point that maximizes posterior density with the Jeffreys measure as the dominating measure [16, 17], eg. for a likelihood $f(x|\theta)$ and prior $p(\theta)$, the parametrization-invariant MAP is given by

$$\theta_{MAP} = \arg \max_{\theta} f(x|\theta) p(\theta) |I(\theta)|^{-1/2}$$

where $I(\theta)$ is the Fisher information associated with $f(x|\theta)$.

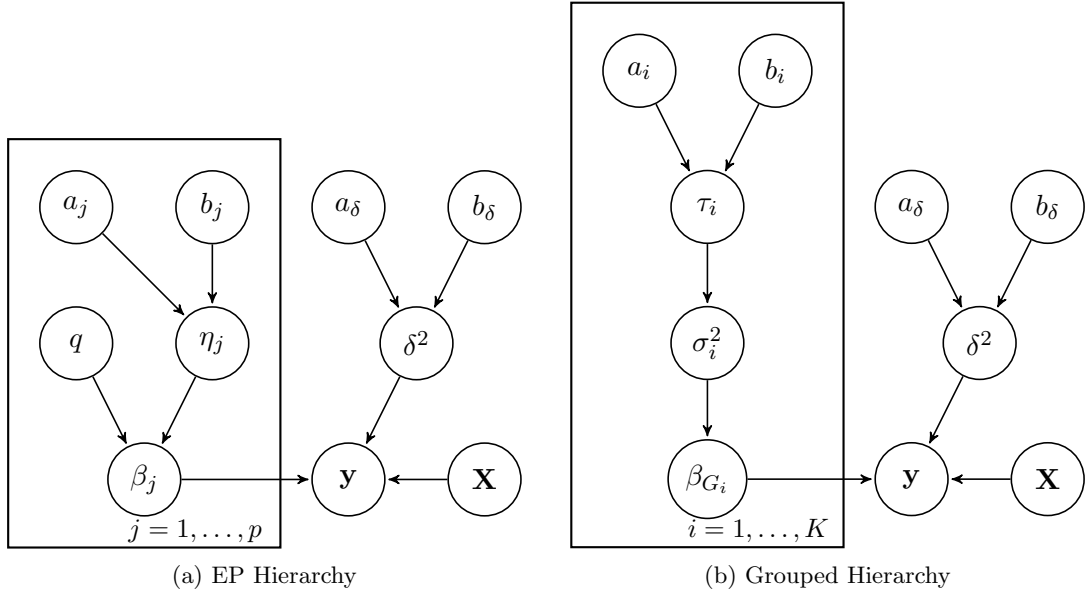


Figure 3: Graphical model representations of the exponential power and the grouped variable hierarchies

An important issue with the MAP estimates obtained from our methodology is that the posterior is multimodal and there is no guarantee that one will obtain the global mode of the posterior as opposed to a local one. However, this is true of almost all non-convex penalized optimization approaches. In [6], a suggestion is to start with high values of b_j and reduce the values of b_j once the algorithm has converged. In principle, this can be done with both a_j and b_j , noting that such an algorithm will still find a local mode of the posterior and this ‘tempering’ of the posterior during optimization can affect which mode is chosen. We do not investigate this further but note that characterization of the modes obtained using such a process is an interesting open question.

3 Related approaches

The proposed approach, either in the hierarchical model or in the estimation step, is closely related to many approaches that have been suggested in the literature. One contribution of this paper is therefore to provide a Bayesian interpretation of existing methods and a flexible framework with which we can incorporate different models.

3.1 Laplacian scale mixture distributions and compressible priors

It has come to our attention that the HAL prior has been proposed independently in both [8] and [9]. In the former, one obtains the same procedure from a majorization-minimization algorithm and in the latter from an EM algorithm. However, our derivation makes explicit the flexibility of the hierarchy and generalizes this prior to exponential power families, situations with grouped variables and positive-definite matrices, making particularly clear strategies for choosing hyperparameters.

3.2 Weakly informative priors and non-convex penalization

With $\beta_j \sim N(0, \sigma_j^2)$ and $\sigma_j^2 \sim IG(a_j, b_j)$ one obtains marginally a t-distribution for $\beta_j | a_j, b_j$. This corresponds to the idea of using weakly informative priors as in [18]. For the case where $\beta_j \sim \text{Laplace}(0, \tau)$ and $\tau \sim IG(a_j, b_j)$, ie. the hierarchical adaptive lasso, we can similarly think of the generalized t-distribution prior on β_j after marginalizing out τ as a weakly informative prior. In fact, one can think of all of the priors proposed using the hierarchical approach in this work as weakly informative.

3.3 Adaptive methods

Within the statistical literature, the closest approach is perhaps the adaptive lasso [3], whose implementation corresponds to a single step of the exponential power family generalization of the HAL with $\beta^{(0)}$ a root- n consistent estimator of β and $b_j \rightarrow 0$. As such, the adaptive lasso estimator can be thought of as taking an initial estimate and returning an estimate with higher posterior density given a logarithmic prior. Our method, on the other hand, finds a local mode of the posterior.

Similarly, the benefit of a polynomial form for the prior density is related to the motivation for the smoothly clipped absolute deviation penalty [2]. Indeed, the penalization induced by the HAL prior grows slowly so that large values of β_j are not unnecessarily biased while remaining continuous and sparse. The methods used to find local linear and local quadratic approximations (LLA and LQA) algorithms of [2, 4] are also closely related, being iteratively reweighted optimization algorithms with a different penalization.

3.4 Iteratively reweighted ℓ_q -minimization

The basic HAL algorithm is clearly similar to the reweighted- ℓ_1 approach proposed in [5], which is identical except that the weights have the form

$$w_j^{(t)} = \frac{\lambda}{\epsilon + |\beta_j^{(t)}|}$$

which corresponds to a limiting case where $a_j \rightarrow \lambda - 1$ and b_j is set to be small.

Similarly, the exponential-family generalization of the HAL algorithm is related to the family of approaches suggested in [6] for the various ℓ_q -penalization norms. As such, the hierarchical model for β gives an interpretation to the methods in the family of iteratively reweighted optimization solutions and, in particular, to the selection of additional parameters ϵ and λ .

3.5 Normal-Exponential-Gamma priors

Our hierarchical prior differs from that suggested in [19] in that an inverse gamma prior is placed on τ_j^2 as opposed to τ_j . This difference in their work results in a posterior for β for which it is difficult to obtain MAP estimates [20], although this problem can be alleviated by novel fast methods for computation of the parabolic cylinder function [21]. The marginal prior

is a member of the generalized hyperbolic family. This difference also appears in [22], although in that work the full posterior is explored using MCMC.

3.6 A note on improper priors

Consider the exponential power density

$$p(\beta_j|\eta_j, q) = \frac{1}{2\eta_j^{1/q}\Gamma(1 + 1/q)} \exp\left(-\frac{|\beta_j|^q}{\eta_j}\right)$$

with the scale-invariant prior on η_j , $p(\eta_j) \propto 1/\eta_j$. The prior on β_j after marginalizing out η_j is then, regardless of q , improper with the form $p(\beta_j|q) \propto 1/|\beta_j|$. Since this is the same prior for $q = 1$, which we know will produce sparse β and for $q = 2$, which is the prior proposed in [23], this explains why the prior in [23] produces sparse results. However, it is worth noting that the posterior for β using this prior is improper with unbounded density at $\beta = \mathbf{0}$.

3.7 Graphical Model

Figure 3 gives graphical models for the hierarchies corresponding to the exponential power (EP) generalization of section 2.3.1 and the adaptive group lasso of section 2.3.2. These models allow us to visualize the flexibility of the framework and the connections with related approaches. Indeed, for $q = 1$ one obtains the hierarchical adaptive lasso or, by setting η_j to be a fixed hyperparameter, the standard lasso. Similarly, for $q = 2$ one obtains hierarchical adaptive ridge regression or standard ridge regression. For the hierarchy with grouped variables, the similarity to the hierarchical adaptive lasso hierarchy is clear, suggesting that application-specific hierarchies could be developed that lead to iteratively reweighted methods.

4 Examples

4.1 Linear regression

In linear regression, the likelihood of \mathbf{y} given X and β is given by

$$p(\mathbf{y}|\mathbf{X}, \beta, \mu, \delta^2) = \frac{1}{(2\pi\delta^2)^{n/2}} \exp\left\{-\frac{1}{2\delta^2}(\tilde{\mathbf{y}}_\mu - \mathbf{X}\beta)^T(\tilde{\mathbf{y}}_\mu - \mathbf{X}\beta)\right\}$$

where $\tilde{\mathbf{y}}_\mu \stackrel{\text{def}}{=} \mathbf{y} - \mu\mathbf{1}_n$. If X is standardized, we have $\mathbf{1}_n^T X = 0$ and so we can put an improper prior on μ with $p(\mu) \propto 1$ and integrate it out so that

$$p(\mathbf{y}|\mathbf{X}, \beta, \delta^2) = \frac{1}{(2\pi\delta^2)^{\frac{n-1}{2}}\sqrt{n}} \exp\left\{-\frac{1}{2\delta^2}(\tilde{\mathbf{y}} - \mathbf{X}\beta)^T(\tilde{\mathbf{y}} - \mathbf{X}\beta)\right\} \quad (3)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_n$.

4.1.1 Fixed δ^2

If δ^2 is fixed, we proceed as expected. Note that in this case, the Jeffreys prior for β is a uniform improper prior so no adjustment needs to be made to make the MAP estimate invariant.

To test the method, we simulated data using $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\delta^2 = 1$ and $X \sim N(0, \Sigma)$ with $\Sigma_{i,j} = 0.5^{|i-j|}$. We then ran 1000 repetitions of the hierarchical adaptive lasso and the standard lasso on this problem with various settings of (a, b) and τ respectively with the results given in Tables 1-2.

Table 1: Results for the LASSO (linear regression, $\delta = 1$)

n	τ	avg. error	% correct	avg. false positives	avg. false negatives
40	0.2	0.4137	9.7	1.808	0.0
40	0.1	0.4817	36.7	0.89	0.0
40	0.02	1.6732	90.0	0.089	0.015
80	0.2	0.2872	2.8	2.519	0.0
80	0.1	0.2931	20.0	1.3510	0.0
80	0.02	0.8169	92.7	0.079	0.0

Table 2: Results for the HAL (linear regression, $\delta = 1$)

n	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
40	(1, 0.1)	0.3118	89.9	0.105	0.0
40	(2, 0.1)	0.3044	98.1	0.019	0.0
40	(2, 0.05)	0.3026	99.6	0.004	0.0
80	(1, 0.1)	0.2191	81.3	0.079	0.0
80	(2, 0.1)	0.2061	96.6	0.034	0.0
80	(2, 0.05)	0.2038	98.8	0.012	0.0

Both methods are capable of giving good results in this setting, which has a high signal-to-noise ratio. However, the reduction in average error is evident for the HAL, owing mainly to less penalization of the selected coefficients. We ran the same experiment but with $\delta = 3$, to test the algorithm with a lower signal-to-noise ratio with the results given in Tables 3-4. Again, the results for the HAL are typically superior to that for the LASSO. However, incorporating prior information leading to less penalization of β_2 and β_5 improves performance drastically. This type of prior information is likely to be necessary when we wish to include variables whose true coefficients are small.

Table 3: Results for the LASSO (linear regression, $\delta = 3$)

n	τ	avg. error	% correct	avg. false positives	avg. false negatives
40	1/6	1.9747	53.5	0.378	0.255
40	0.125	2.3952	43.0	0.207	0.580
40	0.125*	2.2117	93.9	0.005	0.057

* $(\tau_2, \tau_5) = (0.25, 0.25)$

Table 4: Results for the HAL (linear regression, $\delta = 3$)

n	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
40	(2, 0.75)	1.3407	56.2	0.274	0.352
40	(2, 0.1)	1.7198	28.0	0.064	0.831
40	(2, 0.1)*	1.0224	95.9	0.004	0.038

* $(a_2, b_2, a_5, b_5) = (2, 2, 2, 2)$

4.1.2 $\delta^2 \sim IG(a_\delta, b_\delta)$

If we model $\delta^2 \sim IG(a_\delta, b_\delta)$, we can find that MAP estimate associated with the posterior density $p(\boldsymbol{\beta}|\mathbf{y}, X)$, ie. with δ^2 integrated out. To do so, we additionally include δ^2 as a latent variable in the EM algorithm, noting that conditional on $\boldsymbol{\beta}$, δ^2 and $\boldsymbol{\tau}$ are independent. Furthermore, we have $\delta^2|\boldsymbol{\beta}, X, \mathbf{y} \sim IG(a_\delta + (n-1)/2, b_\delta + 1/2(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}))$ For the hierarchical adaptive lasso, we iteratively solve

$$\boldsymbol{\beta}^{(t+1)} = \arg \max_{\boldsymbol{\beta}} -v_j^{(t)} \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) - \sum_{j=1}^p w_j^{(t)} |\beta_j|$$

where

$$v_j^{(t)} = \frac{a_\delta + (n-1)/2}{b_\delta + 1/2(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}^{(t)})} \text{ and } w_j^{(t)} = \frac{a_j + 1}{b_j + |\beta_j^{(t)}|}$$

To test the method, we simulated data using the same as before but letting $\delta^2 \sim IG(a_\delta, b_\delta)$. We then ran 1000 repetitions of the hierarchical adaptive lasso with various settings of $(a_\delta, b_\delta, a, b)$ with the results given in Table 5. There is clearly more difficulty in estimating the coefficients accurately when the variance of the observations is higher and there is again increased performance with good prior information.

Table 5: Results for the HAL (linear regression, random δ^2)

n	(a_δ, b_δ)	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
40	(3, 5)	(2, 0.1)	0.5509	90.5	0.040	0.070
40	(1, 1)	(2, 0.1)	0.7302	79.8	0.058	0.265
40	(1, 4)	(2, 0.2)	1.5046	53.0	0.085	0.742
40	(1, 4)	(2, 0.2)*	1.1865	78.0	0.047	0.318

* $(a_2, b_2, a_5, b_5) = (2, 2, 2, 2)$

4.1.3 Grouped Variable Selection

For grouped variable selection, we use $p = 32$ with groups of size 4. We let $\beta_{1:4} = (3, 1.5, 2, 0.5)'$, $\beta_{9:12} = (6, 3, 4, 1)'$, $\beta_{17:20} = (1.5, 0.75, 1, 0.25)'$ with all other components set to 0. The groupings of variables were given by $G_i = \{4i + k : k \in \{1, 2, 3, 4\}\}$. As with ungrouped variable selection, the hierarchical adaptive version of the group lasso gives lower average errors and has a higher percentage of correct models chosen compared to the standard group lasso.

Table 6: Results for the GLASSO (linear regression, $\delta = 3$)

n	τ	avg. error	% correct	avg. false positives	avg. false negatives
40	1/12	3.4738	65.4	0.580	1.012
40	0.1	3.1407	70.5	0.948	0.432

Table 7: Results for the GHAL (linear regression, $\delta = 3$)

n	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
40	(2, 0.75)	2.1267	89.6	0.328	0.144
40	(2, 0.7)	2.1205	91.1	0.232	0.176

4.2 Logistic regression

In logistic regression with $y_i \in \{-1, 1\}$, one has $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n (1 + \exp(-y_i \boldsymbol{\beta}^T \mathbf{x}_i))^{-1}$ so the log-likelihood is

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = - \sum_{i=1}^n \log(1 + \exp(-y_i \boldsymbol{\beta}^T \mathbf{x}_i)) \quad (4)$$

The Jeffreys prior for this likelihood is given by $p(\boldsymbol{\beta}) \propto |X'VX|^{1/2}$, where V is a diagonal matrix with

$$v_{i,i} = \frac{\exp(-\boldsymbol{\beta}^T \mathbf{x}_i)}{[1 + \exp(-\boldsymbol{\beta}^T \mathbf{x}_i)]^2}$$

As a result, the parametrization-invariant MAP estimate requires us to minimize

$$\boldsymbol{\beta}_{MAP} = \arg \min_{\boldsymbol{\beta}} -\log f(\mathbf{X}|\mathbf{y}, \boldsymbol{\beta}) + \frac{1}{2} \log |X'VX| - \log p(\boldsymbol{\beta})$$

Unfortunately, while $-\frac{1}{2} \log |X'VX|$ is convex, $\frac{1}{2} \log |X'VX|$ is not so the resulting minimization problem is not convex. However, this does not seem to be a serious issue in our examples as the term $\log |X'VX|$ is relatively constant in the regions of high posterior density and so including the MAP correction has little effect on the results.

To test the method, we simulated data using $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $X \sim N(0, \Sigma)$ with $\Sigma_{i,j} = 0.5^{|i-j|}$ as with the linear regression simulations. We then ran 1000 repetitions of the hierarchical adaptive lasso and the standard lasso on this problem with various settings of (a, b) and τ respectively with the results given in Tables 8-9. An interesting result with this example is that for $(a, b) = (2, 0.1)$ except for $(a_2, b_2, a_5, b_5) = (2, 2, 2, 2)$, the HAL gave poor results due to the correlation of the predictors and the relatively high penalization of β_1 . In this case, β_1 was excluded from the model associated with the MAP estimate in every simulation. Using additionally $(a_1, b_1) = (2, 0.5)$ led to a drastic improvement in the results, highlighting the importance the prior hyperparameters can have.

4.3 Gaussian graphical models

The log-likelihood for this model (after standardization) is

$$\log p(X|\Omega) = \frac{n}{2} \log |\Omega| - \frac{n}{2} \text{tr}(S\Omega)$$

Table 8: Results for the LASSO (logistic regression)

n	τ	avg. error	% correct	avg. false positives	avg. false negatives
80	1/7.5	2.8559	62.1	0.387	0.098
80	0.1*	2.7212	93.9	0.008	0.053

* $(\tau_2, \tau_5) = (1, 1)$

Table 9: Results for the HAL (logistic regression)

n	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
80	(2, 0.65)	1.3736	65.4	0.33	0.114
80	(2, 0.1)*	3.2084	0.0	0.00	1.000
80	(2, 0.1) [†]	1.1228	99.2	0.00	0.008

* $(a_2, b_2, a_5, b_5) = (2, 2, 2, 2)$

[†] $(a_1, b_1, a_2, b_2, a_5, b_5) = (2, 0.5, 2, 2, 2, 2)$

where $S = 1/n \sum_{i=1}^n x_i^T \Omega x_i$

Jeffreys prior for this likelihood is given by $p(\Omega) \propto |\Omega|^{(p+1)/2}$ so we can find the parametrization-invariant MAP using

$$\Omega^{(t+1)} = \arg \max_{\Omega} \frac{n-p-1}{2} \log |\Omega| - \frac{n}{2} \text{tr}(S\Omega) - \sum_{i=1}^p \sum_{j=i}^p w_{ij}^{(t)} |\Omega_{ij}|$$

where

$$w_{ij}^{(t)} = \frac{a_{ij} + 1}{b_{ij} + |\Omega_{ij}^{(t)}|}$$

In order for the likelihood with the MAP correction term to be concave, we require $n > p + 1$ since $-\log \det$ is a convex function.

To test the method, we simulated data using

$$\Omega = \begin{pmatrix} 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.3 & 0 & 0.2 & 0 & 0 \\ 0 & 0.2 & 0.3 & 2 & 0 & 0 & 0 & 1.5 \\ 0.5 & 0.8 & 0 & 0 & 1 & 0 & 0.5 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0.5 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0.3 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 & 0 & 0 & 0 & 2 \end{pmatrix}$$

and again used 1000 repetitions of the procedure using the LASSO and the HAL with the results given in Tables 10-11. The HAL clearly has superior performance when using hyperparameters such that the average number of false positives and false negatives are roughly equal.

Table 10: Results for the LASSO (GGM)

n	τ	avg. error	% correct	avg. false positives	avg. false negatives
40	1/45	4.676	23.9	2.789	1.887
40	1/50	4.22	22.3	2.081	2.139

Table 11: Results for the HAL (GGM)

n	(a, b)	avg. error	% correct	avg. false positives	avg. false negatives
40	(1, 0.075)	2.594	65.4	1.304	1.290
40	(2, 0.1)	2.850	57.7	1.343	1.507

5 Discussion

We have proposed a MAP-based variable selection method using a hierarchical prior for β that works reasonably well in practice and brings together a variety of related approaches in the literature. In particular, the estimate itself corresponds to the solution of a non-convex penalized optimization problem, with properties similar to that in [2], ie. estimates of large coefficients tend to be penalized less than in standard ℓ_1 -penalized optimization approaches, while still being sparse and continuous in the data. A possibly more important contribution is the interpretation the method gives for various methods that have been proposed without Bayesian interpretations, in particular for adaptive or one-step methods in the statistics literature and for iteratively reweighted methods in the machine learning and signal processing literatures. This interpretation allows for manipulation of the hierarchy in application-specific ways.

A number of open questions remain when using this class of methodology for variable selection. One is how to resolve the issue of multimodality of the posterior due to the non-concavity of the log of the prior density. Another is assessing the utility of point estimates when there is little guarantee that the model corresponding to the MAP estimate has significant posterior mass from a Bayesian variable selection perspective. In this work, we feel these issues are secondary as the major contribution is in the Bayesian interpretation and generalization of increasingly popular penalized optimization methods amongst practitioners.

Acknowledgments

We would like to thank Mark Schmidt for the public availability and usability of his ℓ_1 -penalized optimization code, which made the implementation of our methods straightforward.

References

- [1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statistical Soc. B*, 58:267–288, 1996.
- [2] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.
- [3] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.
- [4] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533, 2008.

- [5] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.
- [6] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.
- [7] Arthur. P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.
- [8] Pierre Jerome Garrigues. *Sparse Coding Models of Natural Images: Algorithms for Efficient Inference and Learning of Higher-Order Structure*. PhD thesis, University of California, Berkeley, 2009.
- [9] Volkan Cevher. Learning with compressible priors. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 261–269. 2009.
- [10] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.
- [11] Mike West. On scale mixtures of normal distributions. *Biometrika*, 74:646–648, 1987.
- [12] Stephen G. Walker and Eduardo Gutiérrez-Peña. Robustifying Bayesian procedures. In *Bayesian Statistics 6*, pages 685–710. Oxford University Press, New York, 1999.
- [13] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. Technical report, Department of Statistics, University of California, Berkeley, 2006.
- [14] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.
- [15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.
- [16] Ian H. Jermyn. Invariant Bayesian estimation on manifolds. *Annals of Statistics*, 33(2):583–605, 2005.
- [17] Pierre Druilhet and Jean-Michel Marin. Invariant HPD credible sets and MAP estimators. *Bayesian Analysis*, 2(4):681–692, 2007.
- [18] Andrew Gelman, Aleks Jakulin, Maria Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383, 2008.
- [19] Jim E. Griffin and Philip J. Brown. Bayesian adaptive lassos with non-convex penalization. Technical report, IMSAS, University of Kent, 2007.
- [20] Jim E. Griffin and Philip J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.
- [21] Kevin P. Murphy. *Machine Learning: a Probabilistic Approach*. MIT Press. To be published.
- [22] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

- [23] Mario A. T. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159, 2003.